

Beyond Validation

Using programmed diagnostics to
learn about, monitor, and
successfully complete your DH
project

Martin Holmes and Joey Takeda
University of Victoria Endings Project



University
of Victoria

The Endings Project

- How to complete, publish and walk away from your digital edition project...



University
of Victoria

The Endings Project

- How to complete, publish and walk away from your digital edition project...
- ...and have it last for 50 years.



Social Sciences and Humanities
Research Council of Canada

Conseil de recherches en
sciences humaines du Canada

Canada



University
of Victoria

The Projects

- The Map of Early Modern London ↗
- Le Mariage sous l'ancien régime ↗
- The Nxaʔamxcín Dictionary Database
- The Robert Graves Diary ↗
- The Scandinavian-Canadian Studies Journal ↗



Validation is great

- ...especially if you have created a solid, tightly-constrained schema.
- But it's not enough...



Extra Schematron

is a great help. We use Schematron to:

- enforce curly apostrophes in text
- ban quotation marks (use `<quote>` etc.)
- ban leading and trailing spaces in some tags
- enforce correct capitalization
- limit the length of abstracts
- ...and lots more...

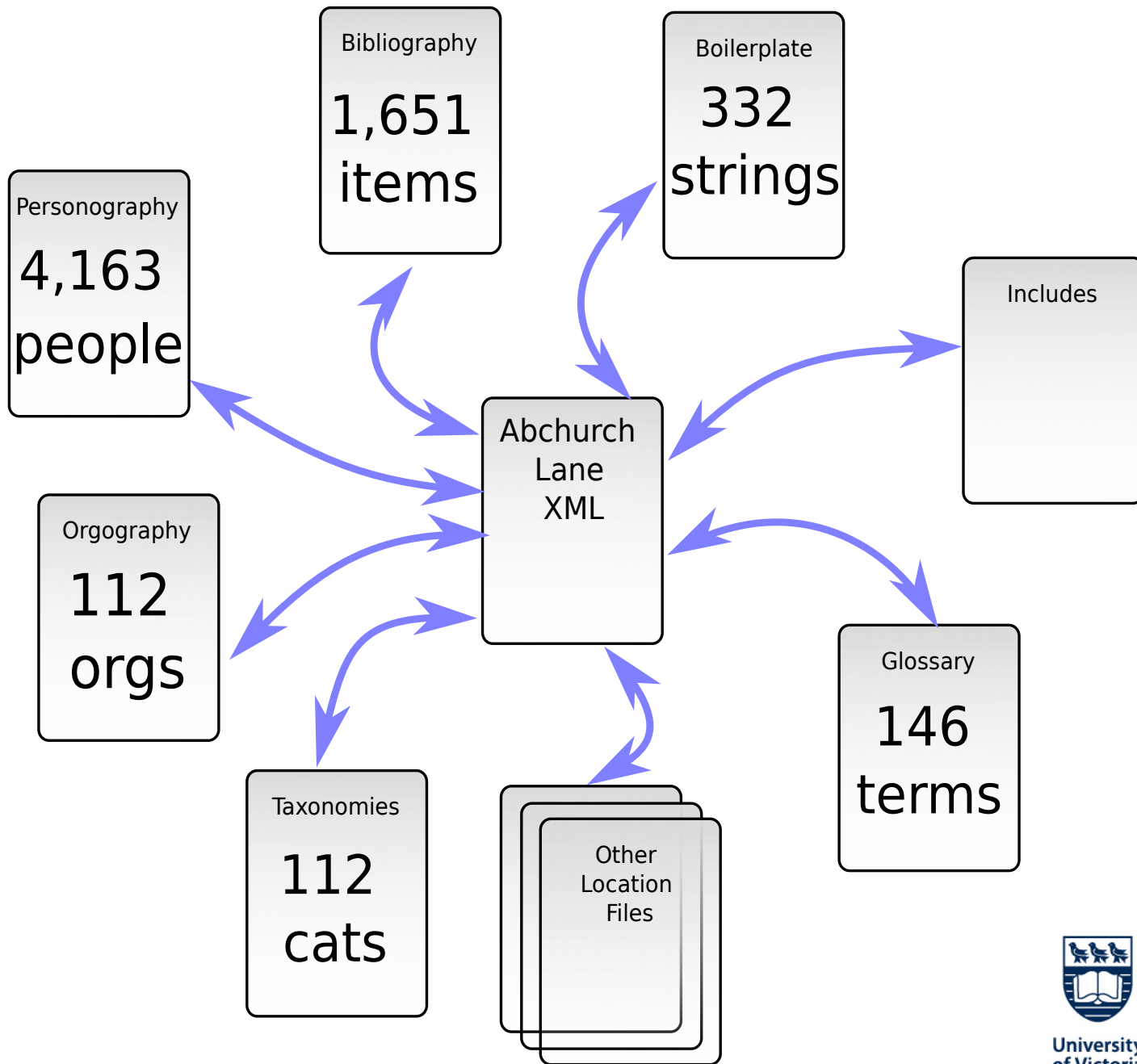


But that's not enough

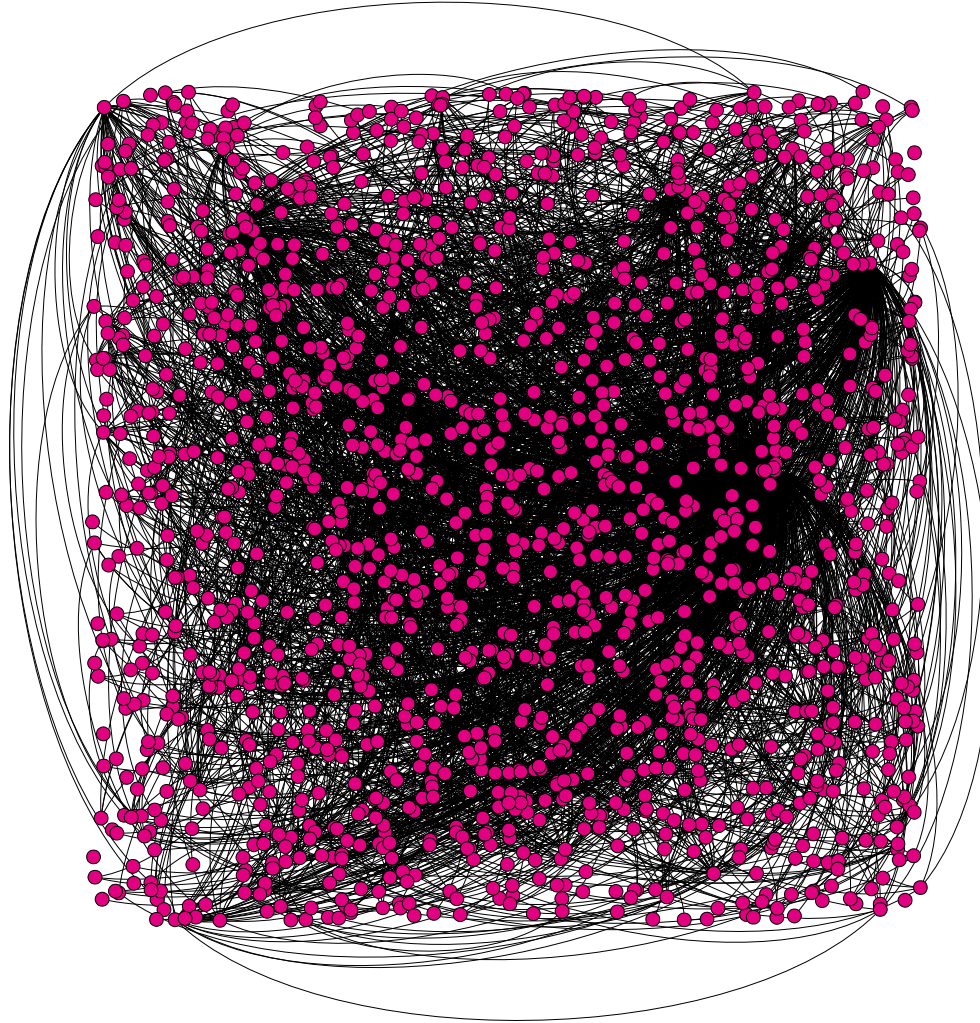
Validation works at the document level, but a digital edition project is typically a large collection of interrelated resources.



University
of Victoria



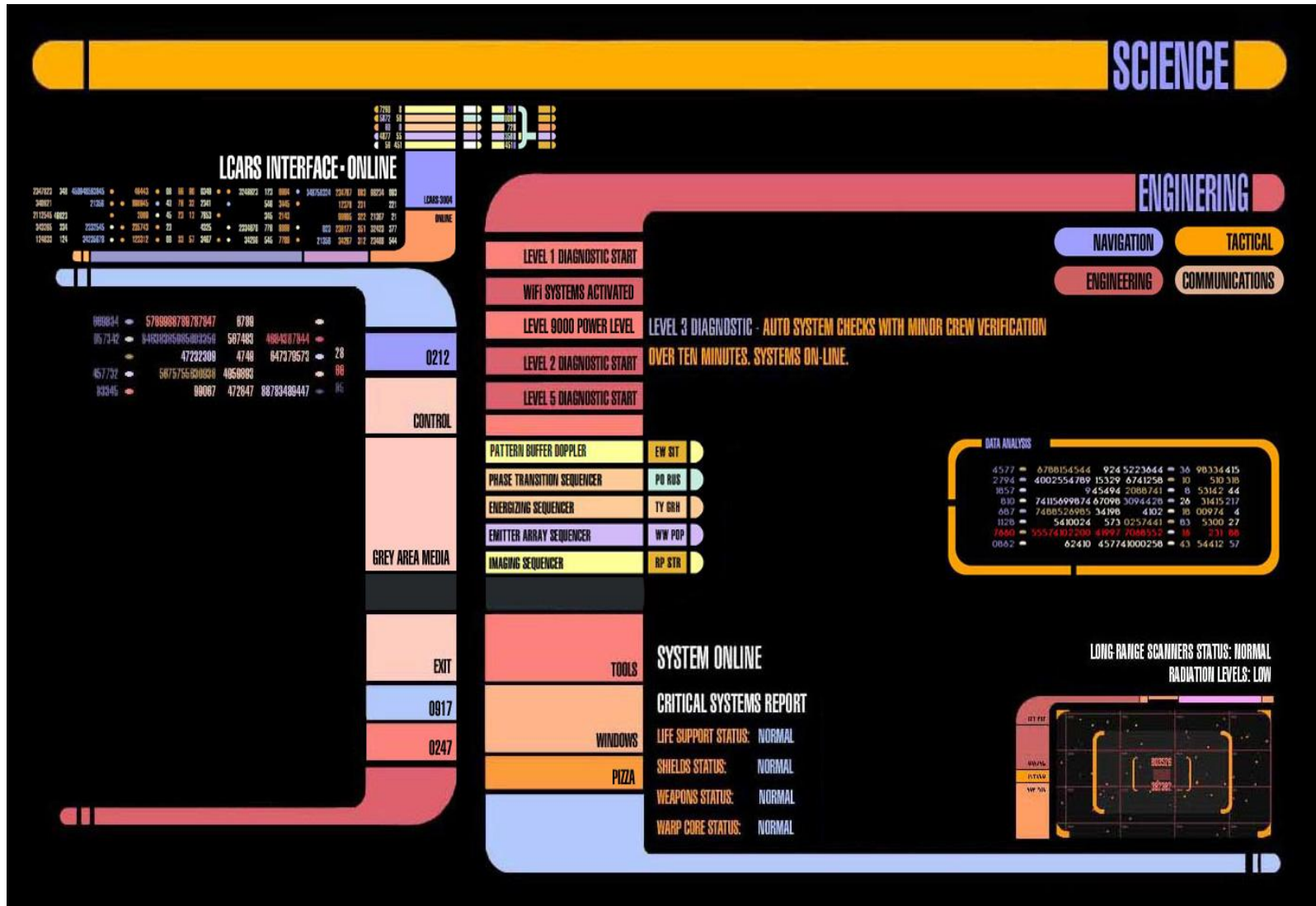
Network graph



University
of Victoria



Diagnostics to the rescue!



COHERENCE and CONSISTENCY

COHERENCE and CONSISTENCY

Project-level checking of referential integrity:

COHERENCE and CONSISTENCY

Project-level checking of referential integrity:

non-existent pointers/targets

COHERENCE and **CONSISTENCY**

Project-level checking of referential integrity:

- non-existent pointers/targets

- duplicate @xml:ids

COHERENCE and CONSISTENCY

Project-level checking of referential integrity:

- non-existent pointers/targets

- duplicate @xml:ids

- entity type errors (persName / placeName / orgName)

DIAGNOSTICS

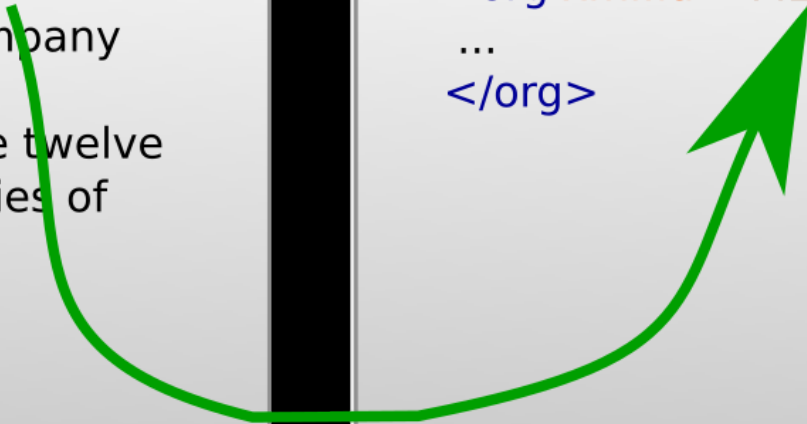
LEVEL 1

PERS1.xml

The
<name type="org"
 ref="mol:MERC3">
 Mercers' Company
 </name>
was one of the twelve
great companies of
London.

ORGS1.xml

<org xml:id="MERC3">
...
</org>



- Every `<name type="org">` points at an `@xml:id` which exists in the project.
- The element pointed at by `<name type="org">` is an `<org>` element in the ORGS1.xml document.
- Every `<name>` element which points at an `<org>` element in ORGS1.xml has `@type="org"`.



- Every `<name type="org">` points at an `@xml:id` which exists in the project.
- The element pointed at by `<name type="org">` is an `<org>` element in the ORGS1.xml document.
- Every `<name>` element which points at an `<org>` element in ORGS1.xml has `@type="org"`.
- Can Schematron do this?



- Every `<name type="org">` points at an `@xml:id` which exists in the project.
- The element pointed at by `<name type="org">` is an `<org>` element in the ORGS1.xml document.
- Every `<name>` element which points at an `<org>` element in ORGS1.xml has `@type="org"`.
- Can Schematron do this?
- Yes, but only for small projects.



DIAGNOSTICS

LEVEL 2

COMPLETENESS

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

Suspiciously similar bibliography entries

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

Suspiciously similar bibliography entries
different names tagged to point to the
same <person>

COMPLETENESS

Generate to-do lists, identify possible errors requiring human judgement:

- Suspiciously similar bibliography entries
- different names tagged to point to the same <person>
- unreachable (unlinked) documents

Possible duplicate <bibl> entries (24) ▼

Explanation

These bibliography entries appear very similar, as measured by a similarity metric, so they are possibly duplicates.

- *Closer to 1 indicates higher similarity.*
- *Closer to 0 indicates lower similarity.*
- <bibl> DEKK2 appears similar to SMAL1 (similarity score 0.925):

Dekker, Thomas. The Shoemaker's Holiday. Ed. R.L. Smallwood and Stanley Wells. Manchester: Manchester UP, 1979. The Revels Plays.

Smallwood, R.L., and Stanley Wells, eds. The Shoemaker's Holiday. By Thomas Dekker. Manchester: Manchester UP, 1979. The Revels Plays.
- <bibl> DEKK13 appears similar to SMUT1 (similarity score 0.90697676):

Dekker, Thomas, Stephen Harrison, Ben Jonson, and Thomas Middleton. The Whole Royal and Magnificent Entertainment of King James through the City of London, 15 March 1604, with the Arches of Triumph. Ed. R. Malcolm Smuts. Thomas Middleton: The Collected Works. Gen. ed. Gary Taylor and John Lavagnino. Oxford: Oxford UP, 2007. 219-79.

Smuts, R. Malcolm, ed. The Whole Royal and Magnificent Entertainment of King James through the City of London, 15 March 1604, with the Arches of Triumph. Thomas Middleton: The Collected Works. Gen. ed. Gary Taylor and John Lavagnino. Oxford: Oxford UP, 2007.

Beware

- Diagnostics can get out of hand.
- The Moses Dictionary Project has gone nuts with **their diagnostics**.
- Possible result: project paralysis.



DIAGNOSTICS

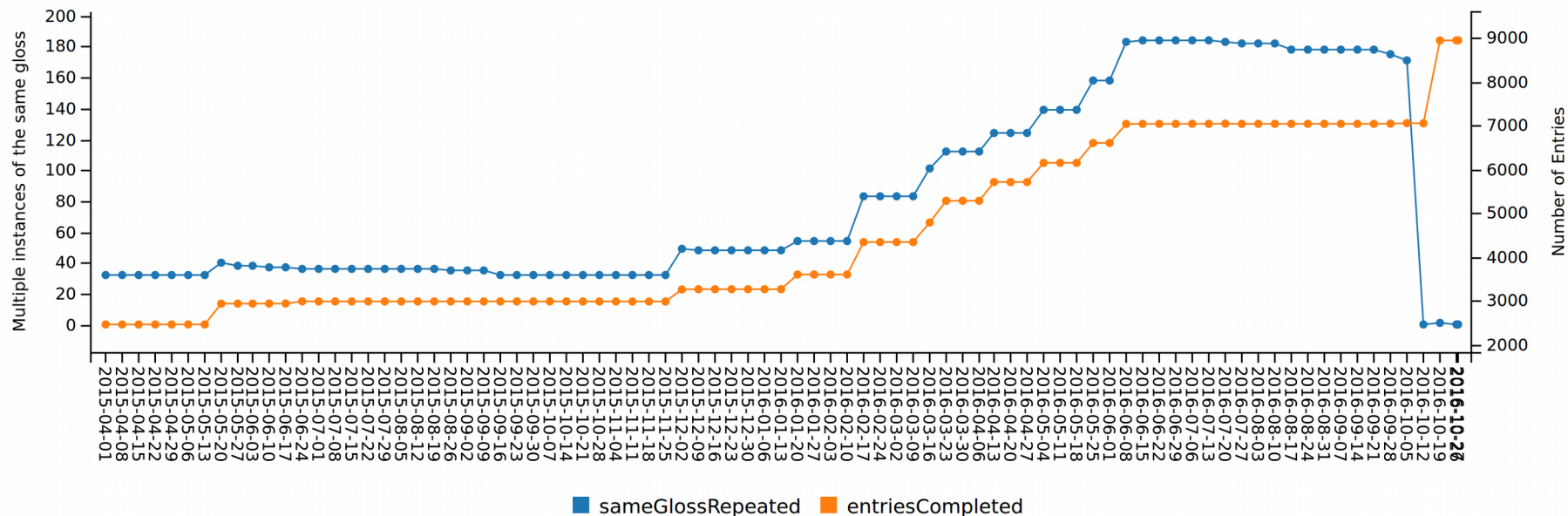
LEVEL 3

PROGRESS AND PREDICTIONS

PROGRESS AND PREDICTIONS

Provide progress analysis: are we hitting milestones?

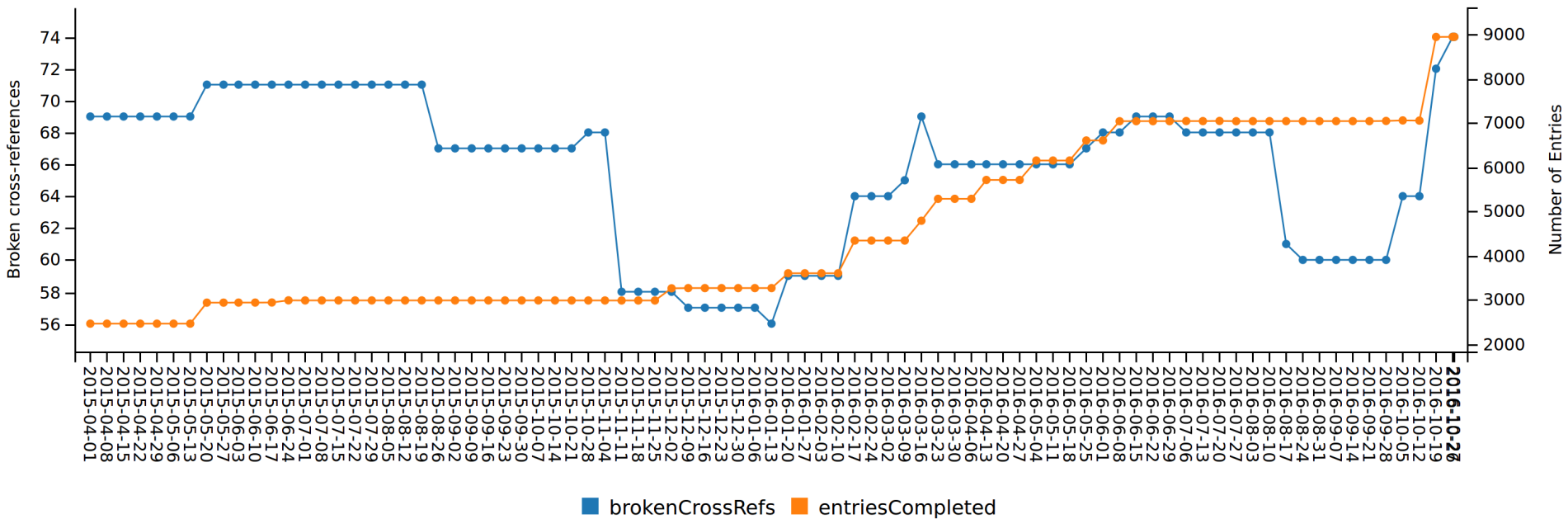
How are we doing?



Duplicate glosses rise with new entries, until an editing blitz fixes them.



University
of Victoria

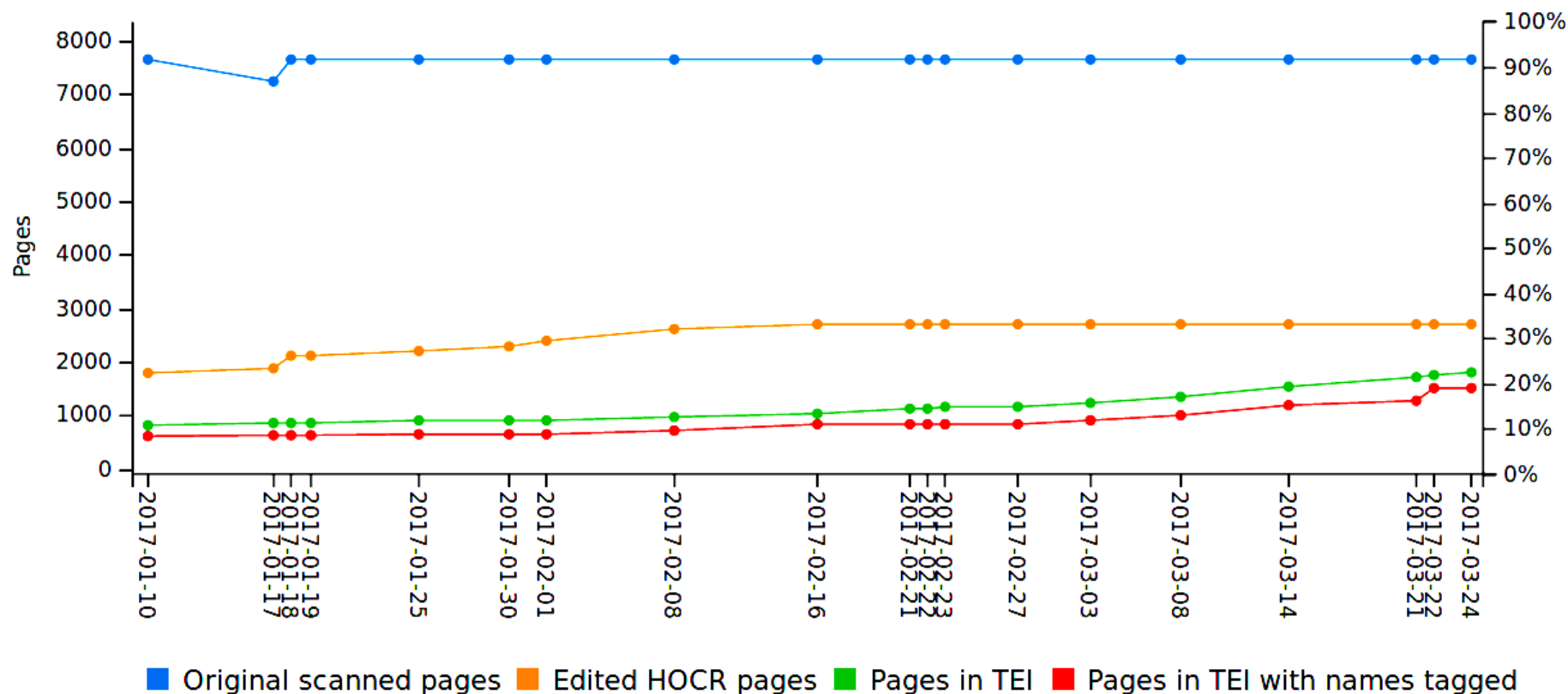


Broken cross-references keep coming back.



University
of Victoria

The Confederation Debates: Progress Chart



Total names tagged so far: 5626

Problematic "unspecified" names: 129

Projected completion dates

- edited HOCR pages: 2017-08-04
- pages in TEI: 2017-11-16
- fully-name-tagged pages in TEI: 2018-01-15



A basic diagnostic toolkit

<https://github.com/projectEndings/diagnostics>



University
of Victoria

- Run it against a folder containing a TEI project.
- It checks that:
 - All pointer attributes within a document point to **@xml:ids** that exist in the document.
 - All pointers to other documents in the collection, or to **@xml:id** attributes in those documents, are correct.
 - All values for the **@xml:lang** attribute are correct.





Project

diagnostics.xpr

- diagnostics.xpr
 - diagnostics
 - images
 - moeml
 - test
 - utilities
 - xsl
 - build.xml
 - buildRelease.xml
 - ids_temp.xml
 - instructions.html
 - README.md



Instructions

Before you do anything, make sure you have a local copy of the diagnostics project. You can either clone it from the GitHub repository, or you can just download a release. You will also need to know the location of your own project's TEI XML files on your computer.

How to run diagnostics on your project in Oxygen

To run diagnostics against a TEI XML encoding project, you will need:

- A project folder which contains all of your TEI XML files. They can be nested in subfolders or organized in any way.
- A single file of any kind in that root folder.

1. In Oxygen, choose **File / Open Project**, and choose the **diagnostics.xpr** file to open the diagnostics project.
2. Open this document in Oxygen

Text Grid Author

/Users/.../projectEndings/diagnostics/instructions.html Learn completed U+0000

Custom Validation

Project

diagnostics.xpr

- diagnostics
 - images
 - moeml
 - test
 - utilities
 - xsl
 - build.xml
 - buildRelease.xml
 - instructions.html
 - README.md

Instructions

Before you do anything, make sure you have a local copy of the diagnostics project. You can either clone it from the GitHub repository, or you can just download a release. You will also need to know the location of your own project's TEI XML files on your computer.

How to run diagnostics on your project in Oxygen

To run diagnostics against a TEI XML encoding project, you will need:

- A project folder which contains all of your TEI XML files. They can be nested in subfolders or organized in any way.
- A single file of any kind in that root folder.

1. In Oxygen, choose **File / Open Project**, and choose the **diagnostics.xpr** file to open the diagnostics project.
2. Open this document in Oxygen

Text Grid Author

U+0000

instructions.html [/Users/joeytakeda/projectEndings/diagnostics/instructions.html] - <oXygen/> XML Editor (Academic use only)

Custom Validation

Project

diagnostics.xpr

- diagnostics
 - images
 - moeml
 - test
 - utilities
 - xsl
 - build.xml
 - buildRelease.xml
 - instructions.html
 - README.md

instructions.html

Choose the directory containing your TEI files

graves

Name	Date Modified
externals	Saturday, March 4, 2017 10:53 PM
graves.xpr	Friday, May 5, 2017 9:40 PM
js	Saturday, March 4, 2017 10:52 PM
notes	Tuesday, July 12, 2016 10:27 PM
presentations	Sunday, March 5, 2017 10:29 PM
products	Saturday, June 17, 2017 1:51 PM
result.xml	Saturday, June 17, 2017 1:50 PM
site	Monday, July 18, 2016 8:43 PM
SVNINFO	Saturday, June 17, 2017 1:49 PM
utilities	Friday, May 5, 2017 8:43 PM
VERSION	Saturday, June 17, 2017 1:49 PM
xml	Friday, May 5, 2017 9:33 PM
xsl	Friday, May 5, 2017 8:43 PM

File Format: All Files

Cancel Choose

```
1 Executing:
2 "/Applications/oxygen/.install4j/jre.bundle/Contents/Home/bin/java" -Dant.home=/Applications/oxygen/tools/ant org.apache.ant.Main -f /Users/joeytakeda/projectEndings/diagnostics/build.xml
3
4 Buildfile: /Users/joeytakeda/projectEndings/diagnostics/build.xml
5
6 getDirectory:
7
```

ANT [runDiagnostics]

/Users/.../projectEndings/diagnostics/instructions.html [runDiagnostics] Transformation in progress. U+0000

Diagnostics for project at /Users/joeytakeda/graves/xml

Statistics ▼

TEI documents	107
Declared xml:ids	3283

Elements used ►

Attributes used ►

Consistency Checks

Bad Internal Links (0) ▼

Explanation

These are links in the project to entities within the projects that do not seem to exist.

None found.

Bad @xml:lang Values (0) ►

Last generated: 2017-08-10T11:50:35.182-04:00



DIAGNOSTICS

PASSED

ALL TESTS SUCCESSFULLY COMPLETED